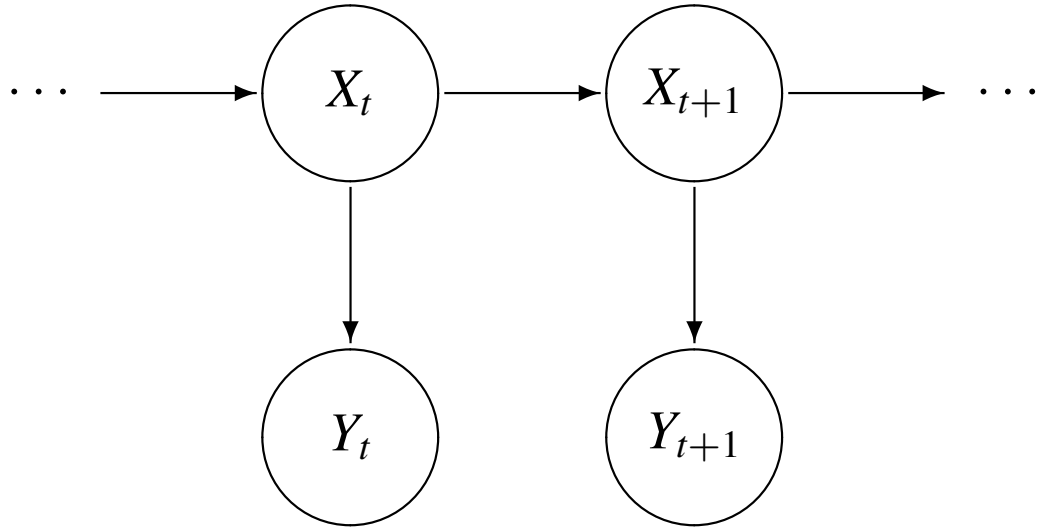
On the Forward Filtering Backward Smoothing particle approximations of the smoothing distribution in general state spaces models

Randal Douc [1], Aurélien Garivier [2] Eric Moulines [2]

[1] Institut Telecom/Telecom SudParis [2] Institut Telecom/Télécom ParisTech

March 2009 / Warwick, MCMC workshop


$\cdots$
X_t
X_{t+1}
$\cdots$
Y_t
Y_{t+1}

Hidden Markov Models

- $\{X_t\}$ is a Markov chain with transition kernel Q and initial distribution χ

$$\mathbb{P}\left(X_{t+1} \in A \,|\, X_t\right) = Q(X_t, A)$$

- The observations $\{Y_t\}$ are conditionally independent given $\{X_t\}$ with conditional density g

$$\mathbb{P}\left(Y_t \in A \,|\, X_t\right) = \int_A g(X_t, y)\lambda(\mathrm{d}y) \ .$$

Motivations

- Statistical inference in general state-space models involves computing the posterior distribution $\phi_{s:s'|t:T}$ of a batch of state variables $X_{s:s'}$ conditioned on a batch of observations $Y_{t:T}$.
- Nonlinearity or non-Gaussianity render analytic solutions intractable. The posterior distribution can be computed in closed form only in very specific cases.
- Problem: How to handle general state and measurement equations without putting strong a priori constraints on the behaviour of the posterior distributions.

Sequential Monte Carlo

- Among these, Sequential Monte Carlo (SMC) methods play a central role.
- SMC methods refer to a class of algorithms designed for approximating a sequence of probability distributions over a sequence of probability spaces by updating recursively in time a set of random particles with associated nonnegative weights.
- These algorithms are all based on selection and mutation and combine sequential importance sampling ideas together with sampling importance resampling.

Smoothing using SMC

- The recursive formulas generating the filtering distribution $\phi_{\chi,T|0:T}$ and the joint smoothing distributions $\phi_{\chi,0:T|0:T}$ are closely related.

$$\phi_{\chi,T|0:T}(f) = \frac{\int \phi_{\chi,0:T-1|T-1}(\mathrm{d}x_{t-1})Q(x_{t-1},\mathrm{d}x_t)g(x_t,y_t)f(x_t)}{\int \phi_{\chi,0:T-1|T-1}(\mathrm{d}x_{t-1})Q(x_{t-1},\mathrm{d}x_t)g(x_t,y_t)}$$

$$\phi_{\chi,0:T|0:T}(f) = \frac{\int \cdots \int \phi_{\chi,0:T-1|T-1}(\mathrm{d}x_{0:t-1})Q(x_{t-1},\mathrm{d}x_t)g(x_t,y_t)f(x_{0:t})}{\int \cdots \int \phi_{\chi,0:T-1|T-1}(\mathrm{d}x_{0:t-1})Q(x_{t-1},\mathrm{d}x_t)g(x_t,y_t)}.$$

- the particle paths and their associated weights is a weighted sample approximating $\phi_{0:T|0:T}$.

Depletion of the particle path

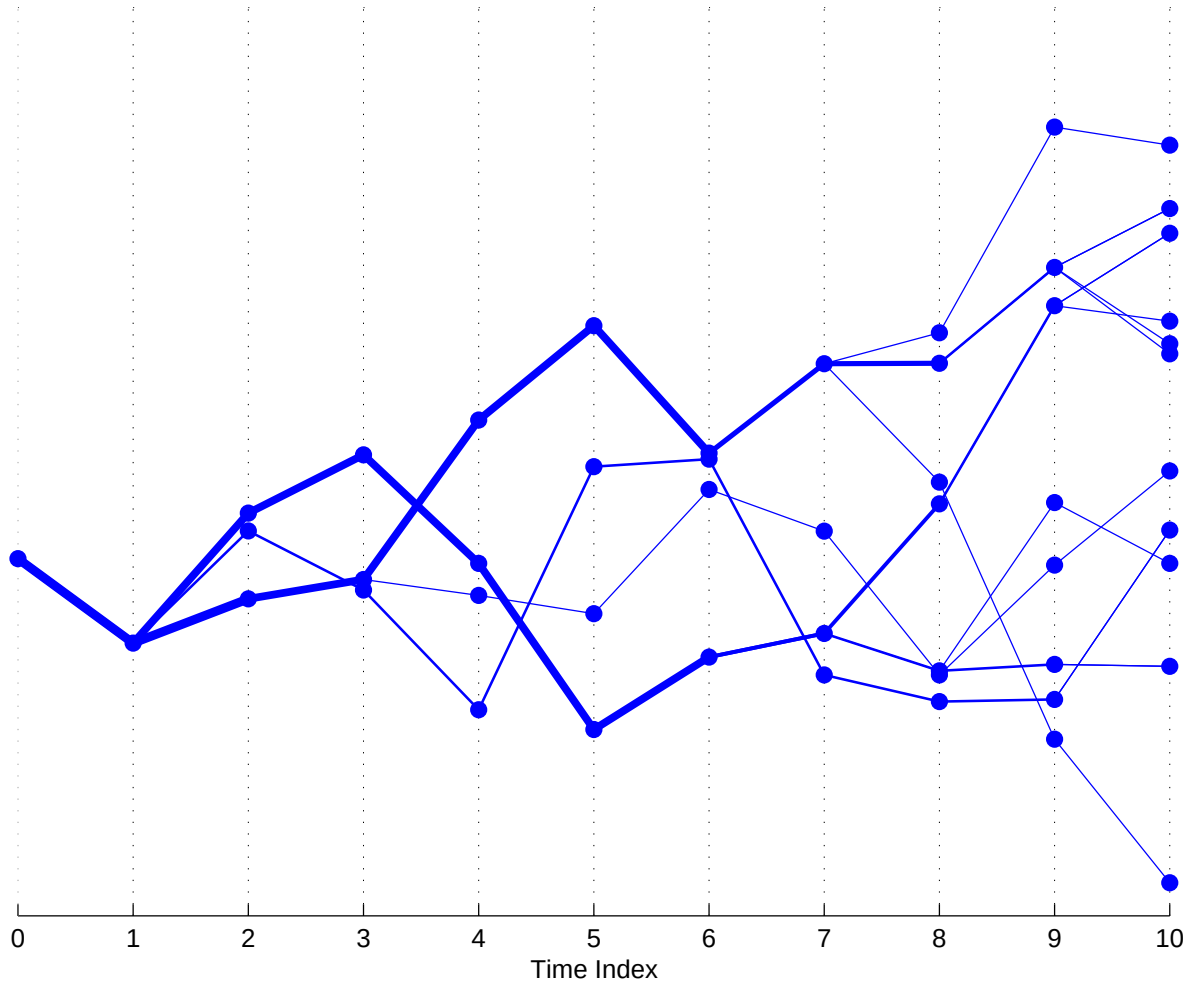


Figure: path trajectories for an AR(1) observed in noise

Problems

- Particle paths can be used successfully for estimating the smoothing joint smoothing distribution for small values of the time horizon T or any marginal smoothing distribution $\phi_{s|0:T}$, with $s \leq T$, when s and T are close;
- However, when T is large or when s and T are remote, the associated particle approximations become inaccurate.
- Calls for alternative solutions.

The backward and forward Markov chains

- Conditionally on the observations $Y_{0:T}$, the state sequence $\{X_s\}_{s\geq 0}$ is a time-inhomogeneous Markov chain:

$$\mathbb{E}\left[f(X_t) \,|\, X_{0:t-1}, Y_{0:T}\right] = \mathbb{E}\left[f(X_t) \,|\, X_{t-1}, Y_{t:T}\right] \ .$$

- This property remains true in the time-reversed direction.

$$\mathbb{E}\left[f(X_t) \,|\, X_{t+1:T}, Y_{0:T}\right] = \mathbb{E}\left[f(X_t) \,|\, X_{t+1}, Y_{0:t}\right] \ .$$

- Given T and an initial distribution χ, the backward kernel $\mathrm{B}_{\chi,s}(x_{s+1}, \cdot)$ is defined as

$$\mathrm{b}_{\chi,t}(x_{t+1}, x) \stackrel{\text{def}}{=} \frac{\phi_{\chi,t|t}(x) q(x, x_{t+1})}{\int \phi_{\chi,t|t}(x) q(x, x_{t+1}) \, \mathrm{d}x} \ .$$

Joint smoothing distribution

- The joint smoothing distribution is

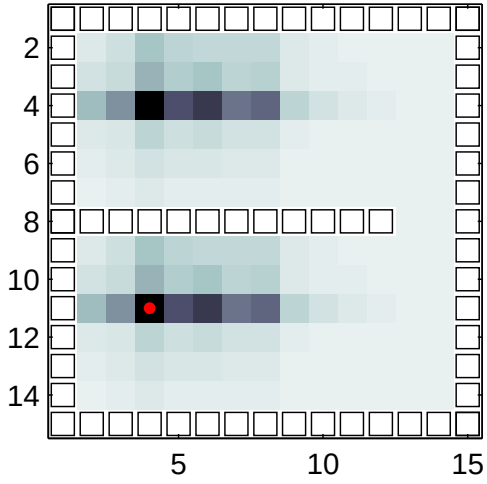
$$\begin{aligned}\phi_{\chi,s:T|T}(f) &= \mathbb{E}_{\chi}\left[f(X_{s:T}) \mid Y_{0:T}\right] \\ &= \int \cdots \int f(x_{s:T})\, \mathrm{B}_{\chi,s}(x_{s+1}, \mathrm{d}x_s)\, \phi_{\chi,s+1:T|T}(\mathrm{d}x_{s+1:T}) \ .\end{aligned}$$

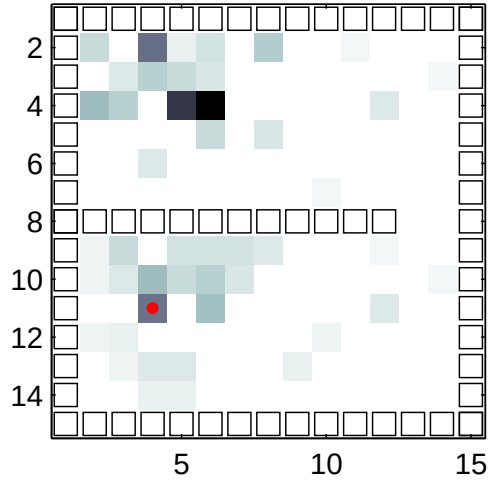
  where $\phi_{\chi,T:T|T} = \phi_{\chi,T|T}$ is the filtering distribution at time T.
- If f depends on the first component x_s only, the marginal smoothing distribution is:

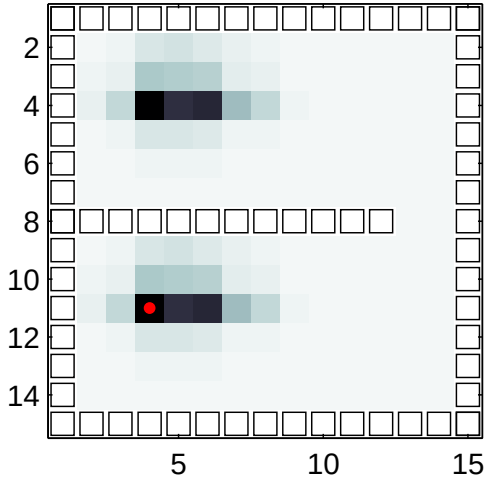
$$\phi_{\chi,s|T}(f) = \iint f(x_s)\, \mathrm{B}_{\chi,s}(x_{s+1}, \mathrm{d}x_s)\, \phi_{\chi,s+1|T}(\mathrm{d}x_{s+1}) \ .$$

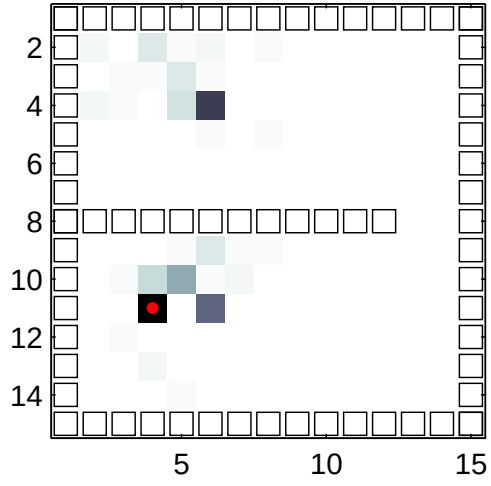
The FFBS algorithm

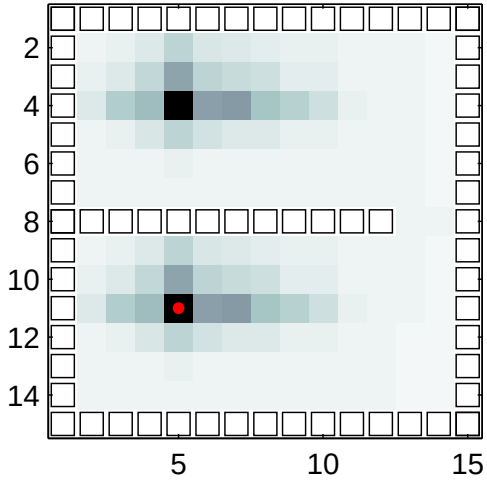
- The FFBS shares many similarities with the forward-backward (Baum-Welsh) algorithm in finite state-space HMM. This is a two-passes algorithm.
- The particle filter is executed, while storing the weighted sample $\left\{\left(\xi_t^i, \omega_t^i\right)\right\}_{i=1}^N$, $1 \leq t \leq T$;
- secondly, starting with the particle approximation of $\phi_{\chi,t|t}$, $t = 0, \ldots, T$, the importance weights are recursively updated backward in time by combining
  1. particle estimates of the fixed interval smoothing distribution $\phi_{\chi,s+1:T|T}$
  2. the filtering distribution estimate $\phi_{\chi,s|s}$.

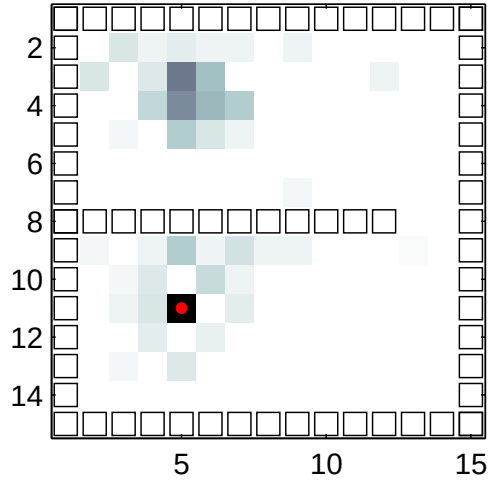

2
4
6
8
10
12
14
5
10
15

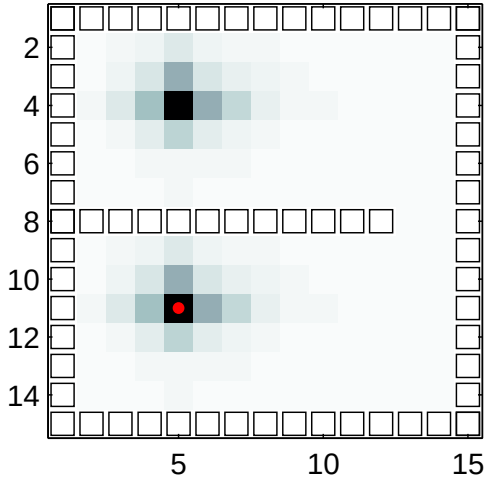

2
4
6
8
10
12
14
5
10
15

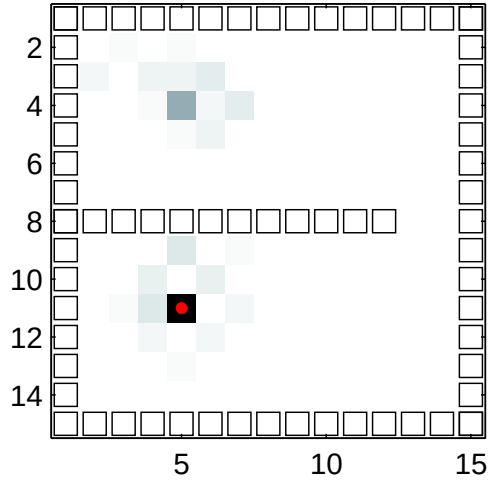

2
4
6
8
10
12
14
5
10
15

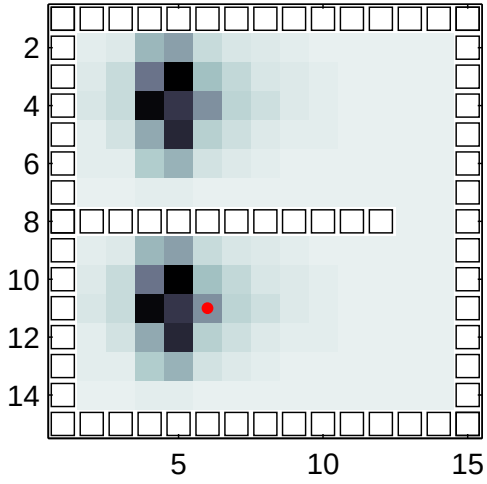

2
4
6
8
10
12
14
5
10
15

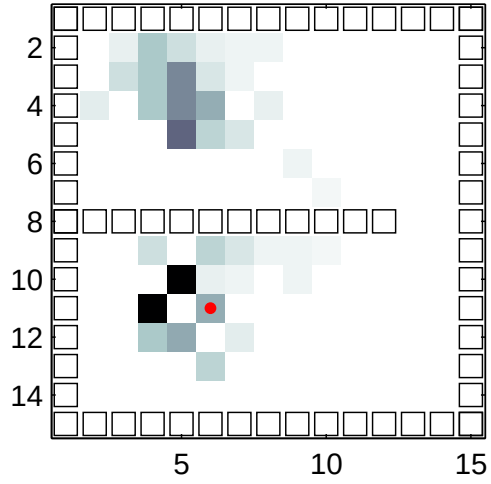

2
4
6
8
10
12
14
5
10
15

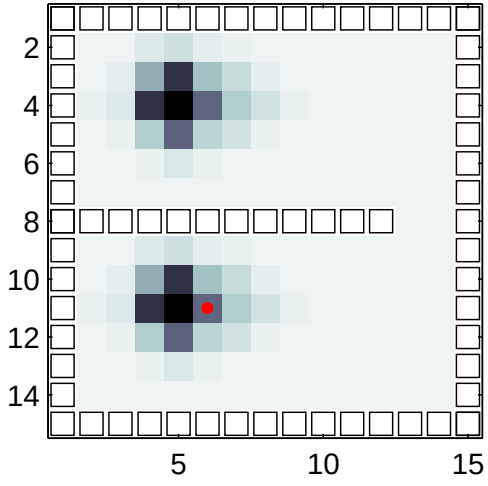

2
4
6
8
10
12
14
5
10
15

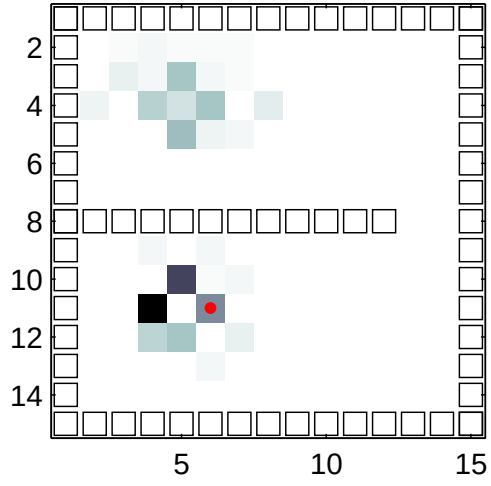

2
4
6
8
10
12
14
5
10
15

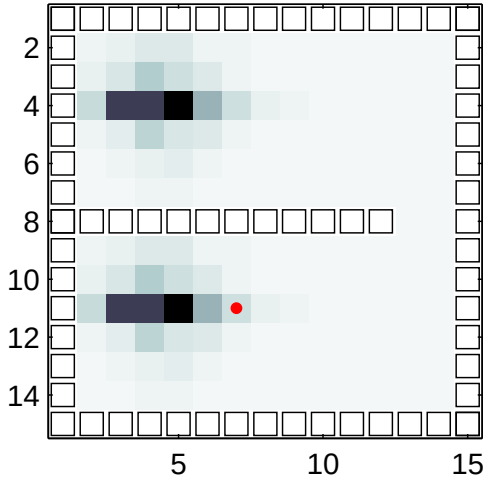

2
4
6
8
10
12
14
5
10
15

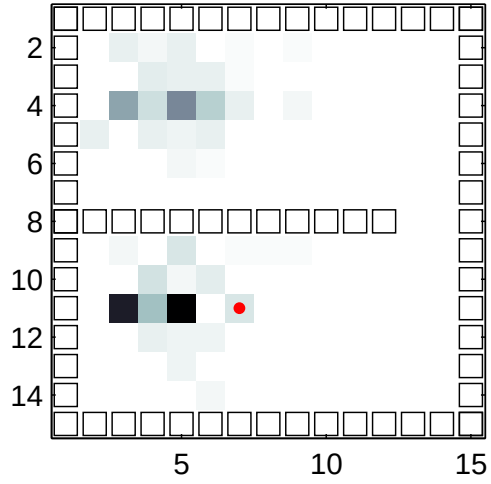

2
4
6
8
10
12
14
5
10
15

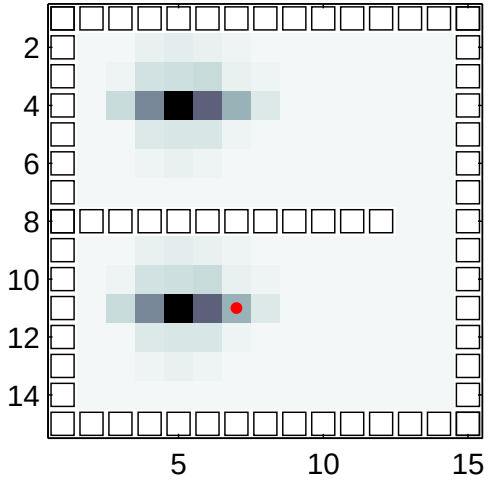

2
4
6
8
10
12
14
5
10
15

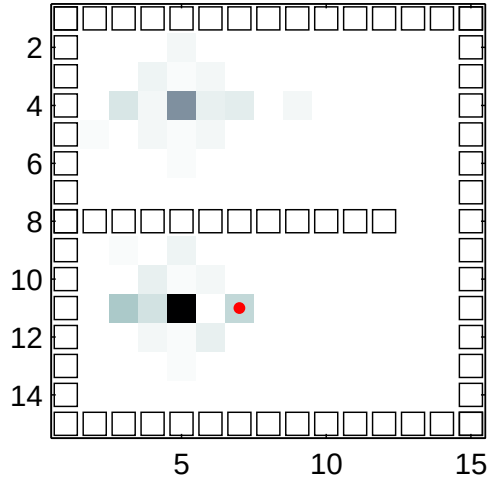

2
4
6
8
10
12
14
5
10
15

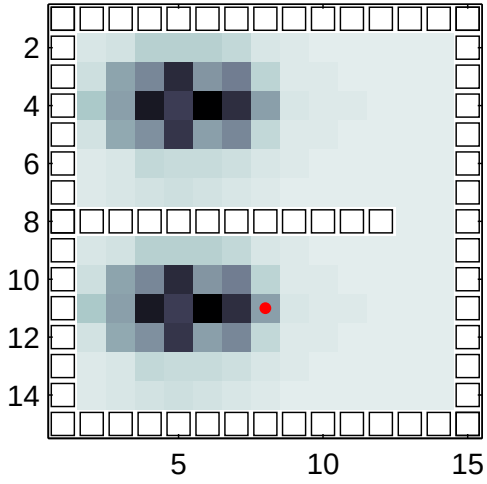

2
4
6
8
10
12
14
5
10
15

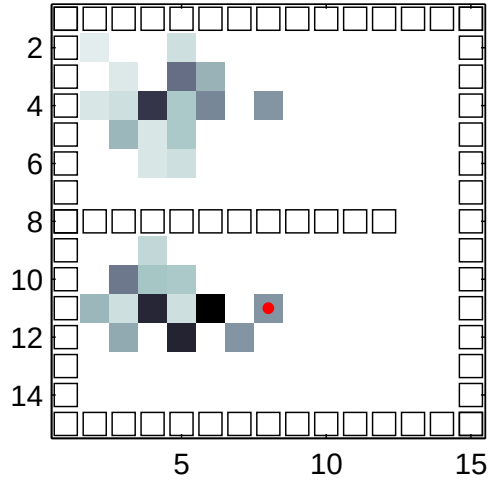

2
4
6
8
10
12
14
5
10
15

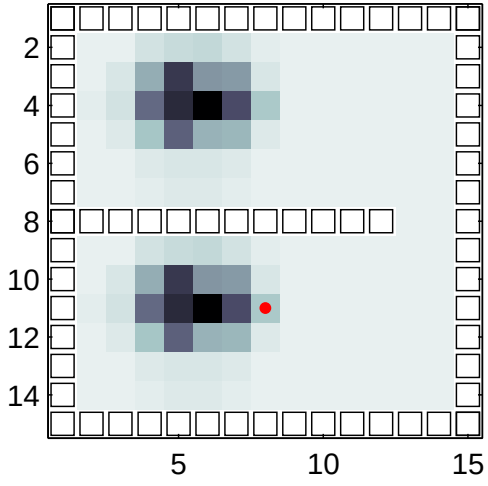

2
4
6
8
10
12
14
5
10
15

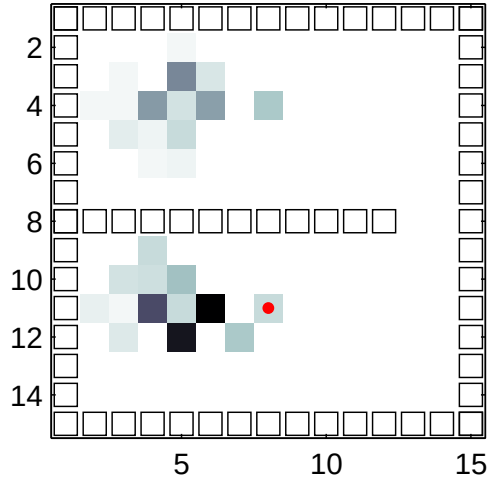

2
4
6
8
10
12
14
5
10
15


2
4
6
8
10
12
14
5
10
15


2
4
6
8
10
12
14
5
10
15


2
4
6
8
10
12
14
5
10
15


2
4
6
8
10
12
14
5
10
15


2
4
6
8
10
12
14
5
10
15

Particle Approximation of the Joint Smoothing Distribution

- For $1 \leq s \leq t \leq T$, define $\xi_{s:t}^{i_{s:t}} \stackrel{\text{def}}{=} (\xi_s^{i_s}, \ldots, \xi_t^{i_t})$, the set of all possible particle paths
- The FFBS approximation of the joint smoothing distribution is

$$\hat{\phi}_{\chi,s:T|T}(\mathrm{d}x_{s:T}) \propto \sum_{j_{s:T}=1}^{N} \omega_{s|T}^{j_{s:T}} \delta_{\xi_{s:T}^{j_{s:T}}}(\mathrm{d}x_{s:T}) \ ,$$

- The support of the joint smoothing distribution is the set of all possible paths... this is why the depletion is expected to be less extreme.... still, the support of the smoothing distribution is selected in the forward pass !

Particle Approximation of the Joint Smoothing Distribution

- Beware: The number of such paths grows exponentially with the horizon $T - s$... of course, this is not implementable except if one is interested in estimating "fixed" dimensional marginal distribution
- The theory is however more transparent (and in fact the proofs essentially requires) to proceed with the full joint distribution...

The FFBS algorithm

- Consider the approximation of the joint smoothing distribution:

$$\hat{\phi}_{\chi,s+1:T|T}(\mathrm{d}x_{s+1:T}) \propto \sum_{j_{s+1:T}=1}^{N} \omega_{s+1|T}^{j_{s+1:T}} \delta_{\xi_{s+1:T}^{j_{s+1:T}}}(\mathrm{d}x_{s+1:T}) \ ,$$

- Approximate the backward smoothing kernel by

$$\hat{\mathrm{B}}_{\chi,s}(x_{s+1}, \mathrm{d}x_s) = \sum_{i=1}^{N} \frac{\omega_s^i q(\xi_s^i, x_{s+1})}{\sum_{\ell=1}^{N} \omega_s^\ell q(\xi_s^\ell, x_{s+1})} \delta_{\xi_s^i}(\mathrm{d}x_s)$$

- Substituting the particle approximations of the backward kernel and the joint smoothing distribution into the backward recursion

$$\phi_{\chi,s:T|T}(f) = \int \cdots \int f(x_{s:T}) \, \mathrm{B}_{\chi,s}(x_{s+1}, \mathrm{d}x_s) \, \phi_{\chi,s+1:T|T}(\mathrm{d}x_{s+1:T})$$

yields to the updating rule for the weights :

$$\omega_{s|T}^{j_{s:T}} = \frac{\omega_s^{j_s} q(\xi_s^{j_s}, \xi_{s+1}^{j_{s+1}})}{\sum_{\ell=1}^{N} \omega_s^\ell q(\xi_s^\ell, \xi_{s+1}^{j_{s+1}})} \omega_{s+1|T}^{j_{s+1:T}} \ .$$

The FFBS as an importance sampling estimator

- The support of this particle estimator are the set of the N^{T-s+1} possible particle paths $\{\xi_{s:T}^{j_{s:T}}\}$.
- The importance weight of these path particles is computed as if the path particle $\xi_{s:T}^{j_{s:T}}$ were simulated by drawing forward in time, for $s < t \leq T$, $\xi_t^{j_t}$ in the set $\{\xi_t^i\}_{i=1}^N$ conditionally independently from the distribution

$$\Omega_{t-1}^{-1} \sum_{\ell=1}^{N} \omega_{t-1}^{\ell} q(\xi_{t-1}^{\ell}, \cdot) \ , \quad i = 1, \ldots, N \ ,$$

which approximates the predictive distribution $\phi_{\chi,t|t-1}$.

The FFBS as an importance sampling estimator

- Beware: The particle paths are not independent ! this approximation would be approximately correct for a finite block of particles selected randomly, using propagation of chaos property.
- This is why standard results on importance sampling estimators cannot be applied to that context.
- This calls for the derivation of specific results.

Questions

- Asymptotic In which sense the FFBS estimator is consistent ? asymptotically normal ?
- Non-asymptotic Can we derive non-asymptotic exponential deviation bounds ?
- Time Uniform results Provided that the kernel Q is properly mixing, does the deviation bound may be shown to be bounded in the long run ? Does the asymptotic variance remain bounded ?
- Better algorithms ?

We derive all these results in the case where the auxiliary particle filter is used in the forward pass.

Auxiliary Particle Filter

- Sample $\{(I_s^i, \xi_s^i)\}_{i=1}^N$ from the proposal distribution

$$\pi_{s|s}(i, x_s) \propto \omega_{s-1}^i \vartheta_s(\xi_{s-1}^i) p_s(\xi_{s-1}^i, x_s) \ ,$$

  where $\{\vartheta_s(\xi_{s-1}^i)\}_{i=1}^N$ are the adjustment multiplier weights and p_s is the proposal transition density function.
- Each draw (I_s^i, ξ_s^i) is assigned to the weight

$$\omega_s^i \stackrel{\text{def}}{=} \frac{q(\xi_{s-1}^{I_s^i}, \xi_s^i) g_s(\xi_s^i)}{\vartheta_s(\xi_{s-1}^{I_s^i}) p_s(\xi_{s-1}^{I_s^i}, \xi_s^i)} \ ,$$

- $\left\{(\xi_s^i, \omega_s^i)\right\}_{i=1}^N$ approximates the target distribution $\phi_{\chi,s|s}$.

Assumptions

- There exist two constants $0 < \sigma_- \leq \sigma_+ < \infty$, such that, for any $(x, x') \in \mathsf{X} \times \mathsf{X}$,
$$\sigma_- \leq q(x, x') \leq \sigma_+ \ .$$
- In addition, there exists a constant $c_- > 0$ such that, $\int \chi(\mathrm{d}x_0) g_0(x_0) \geq c_-$ and for all $t \geq 1$,
$$\inf_{x \in \mathsf{X}} \int q(x, x') g_t(x') \mathrm{d}x' \geq c_- > 0 \ .$$

Note that we do not assume that the likelihood is lower bounded, which is a less stringent assumption than usually done... the proofs are more tricky.

Forward and backward stability

$$\ell_{s,t}(x_s, x_t) \stackrel{\text{def}}{=} \int \cdots \int q(x_s, x_{s+1}) g_{s+1}(x_{s+1}) \prod_{u=s+1}^{t-1} q(x_u, x_{u+1}) g_{u+1}(x_{u+1}) \mathrm{d}x_{s+1:t-1}$$

Proposition

Under the strong ergodicity condition, for all χ, χ', $s \leq t$ and any bounded measurable functions h,

$$\left| \frac{\int\int \chi(\mathrm{d}x_s) \ell_{s,t}(x_s, x_t) h(x_t) \mathrm{d}x_t}{\int\int \chi(\mathrm{d}x_s) \ell_{s,t}(x_s, x_t) \mathrm{d}x_t} - \frac{\int\int \chi'(\mathrm{d}x_s) \ell_{s,t}(x_s, x_t) h(x_t) \mathrm{d}x_t}{\int\int \chi'(\mathrm{d}x_s) \ell_{s,t}(x_s, x_t) \mathrm{d}x_t} \right| \leq \rho^{t-s} \operatorname{osc}(h) \ ,$$

where $\rho \stackrel{\text{def}}{=} 1 - \sigma_-/\sigma_+$. For any bounded non-negative measurable functions f and f',

$$\left| \frac{\int\int \chi(\mathrm{d}x_s) h(x_s) \ell_{s,t}(x_s, x_t) f(x_t) \mathrm{d}x_t}{\int\int \chi(\mathrm{d}x_s) \ell_{s,t}(x_s, x_t) f(x_t) \mathrm{d}x_t} - \frac{\int\int \chi(\mathrm{d}x_s) h(x_s) \ell_{s,t}(x_s, x_t) f'(x_t) \mathrm{d}x_t}{\int\int \chi(\mathrm{d}x_s) \ell_{s,t}(x_s, x_t) f'(x_t) \mathrm{d}x_t} \right| \leq \rho^{t-s} \mathrm{o}$$

Time-Uniform exponential inequality

Proposition

Under the strong mixing assumption, the filtering distribution satisfies a time-uniform exponential deviation inequality, i.e. there exist constants B and C such that, for all integers N and $t \geq 0$, all measurable functions h and all $\epsilon > 0$,

$$\mathbb{P}\left[\left|N^{-1}\sum_{i=1}^{N}\omega_t^i h(\xi_t^i) - \frac{\phi_{\chi,t|t-1}(g_t h)}{\phi_{\chi,t-1|t-1}(\vartheta_t)}\right| \geq \epsilon\right] \leq B\mathrm{e}^{-CN\epsilon^2/|h|_\infty^2} ,$$

$$\mathbb{P}\left[\left|\hat{\phi}_{\chi,t|t}(h) - \phi_{\chi,t|t}(h)\right| \geq \epsilon\right] \leq B\mathrm{e}^{-CN\epsilon^2/\operatorname{osc}^2(h)} .$$

A glimpse at the proof

W.l.o.g., assume that $\phi_{\chi,t|t}(h) = 0$ and decompose as $\hat{\phi}_{\chi,t|t}(h)$

$$\hat{\phi}_{\chi,t|t}(h) = \sum_{s=1}^{t} \left(\frac{B_{s,t}(h)}{B_{s,t}(\mathbf{1})} - \frac{B_{s-1,t}(h)}{B_{s-1,t}(\mathbf{1})} \right) + \frac{B_{0,t}(h)}{B_{0,t}(\mathbf{1})} ,$$

where

$$B_{s,t}(h) = N^{-1} \sum_{i=1}^{N} \omega_s^i \frac{L_{s,t}(\xi_s^i, h)}{|L_{s,t}(\cdot, \mathbf{1})|_\infty} .$$

and derive an exponential bound for all these terms.

- the stability of the filter allows to bound the oscillation norm of $L_{s,t}(\xi_s^i, h)/L_{s,t}(\xi_s^i, \mathbf{1})$ by a quantity proportional to $\rho^{t-s} \operatorname{osc}(h)$
- $L_{s,t}(\xi_s^i, \mathbf{1}) / |L_{s,t}(\cdot, \mathbf{1})|_\infty$ is bounded.
- The proof is then based on the Hoeffding inequality... but is tricky because the denominator (normalization) is random! Requires a special formulation which uses that $L_{s,t}(\xi_s^i, h)/L_{s,t}(\xi_s^i, \mathbf{1})$ is bounded.

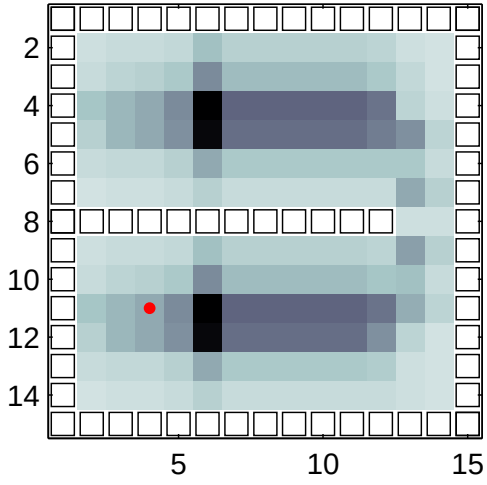
Time-Uniform deviation: marginal smoothing

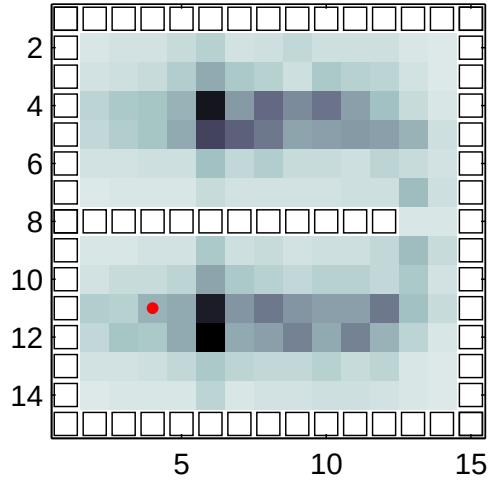
Theorem

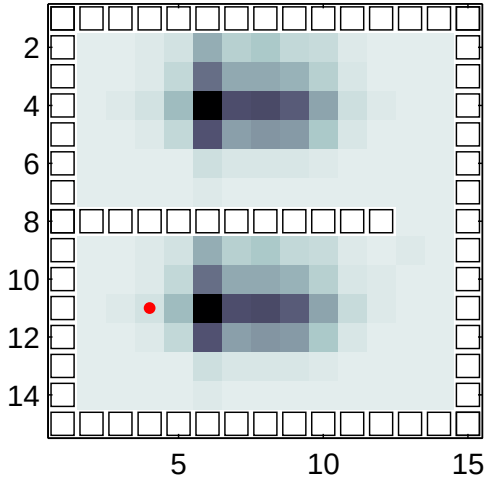
Assume the mixing condition. Then, there exist constants $0 \leq B,\ C < \infty$ *such that for all integers* N, s, *and* T, $s \leq T$, *all* $\epsilon > 0$,

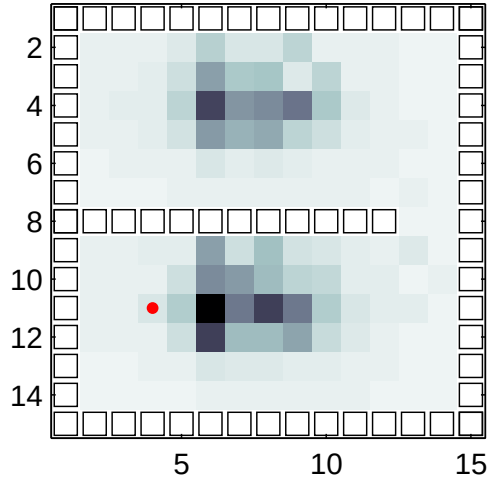
$$\mathbb{P}\left[\left|\hat{\phi}_{\chi,s|T}(h) - \phi_{\chi,s|T}(h)\right| \geq \epsilon\right] \leq B\mathrm{e}^{-CN\epsilon^2 / \operatorname{osc}^2(h)} .$$

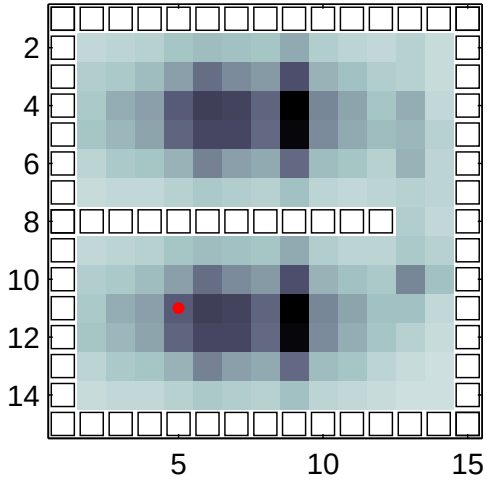
The error does not build up (contrarily to what was initially thought).

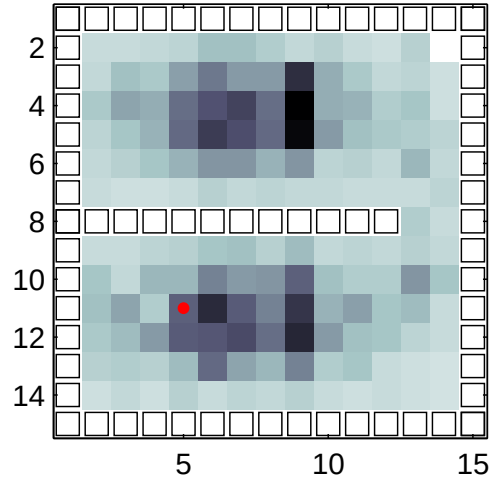

2
4
6
8
10
12
14
5
10
15

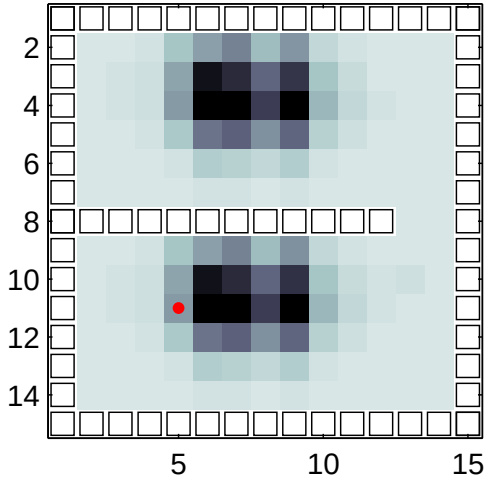

2
4
6
8
10
12
14
5
10
15

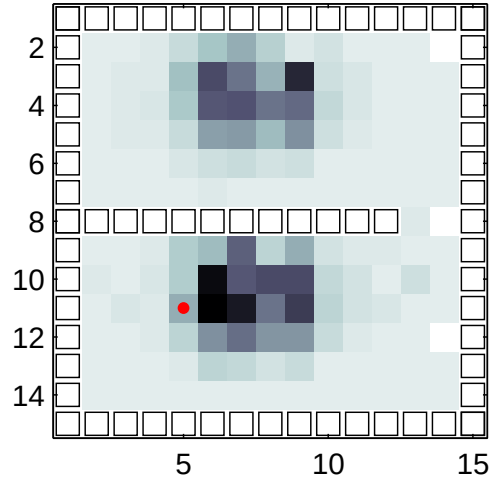

2
4
6
8
10
12
14
5
10
15

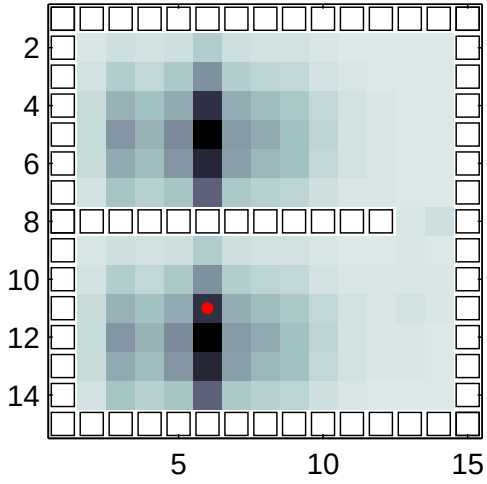

2
4
6
8
10
12
14
5
10
15

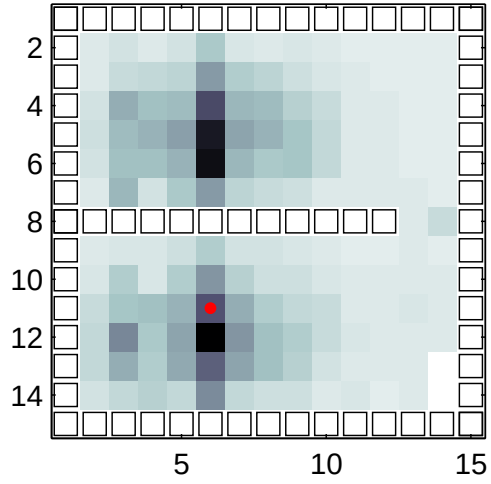

2
4
6
8
10
12
14
5
10
15

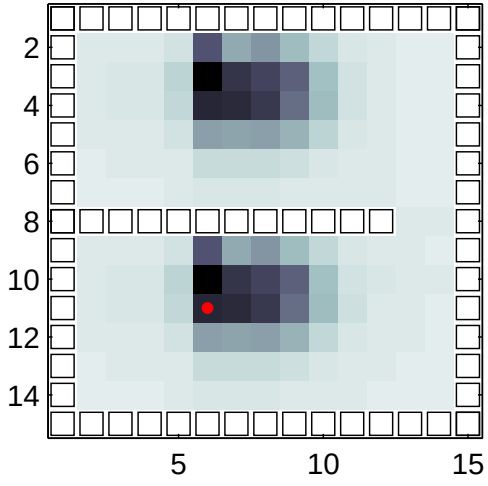

2
4
6
8
10
12
14
5
10
15

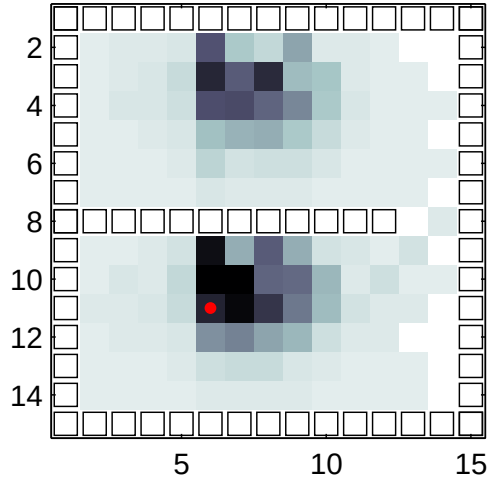

2
4
6
8
10
12
14
5
10
15

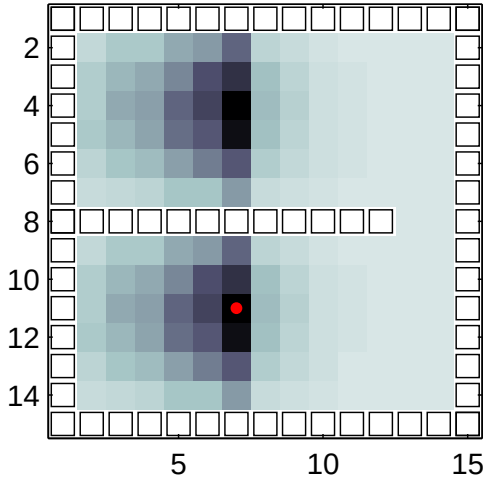

2
4
6
8
10
12
14
5
10
15

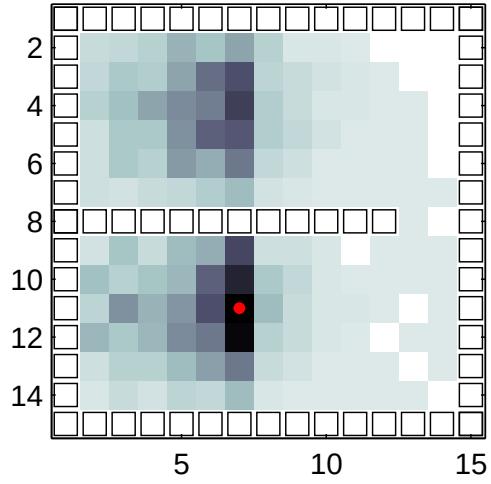

2
4
6
8
10
12
14
5
10
15

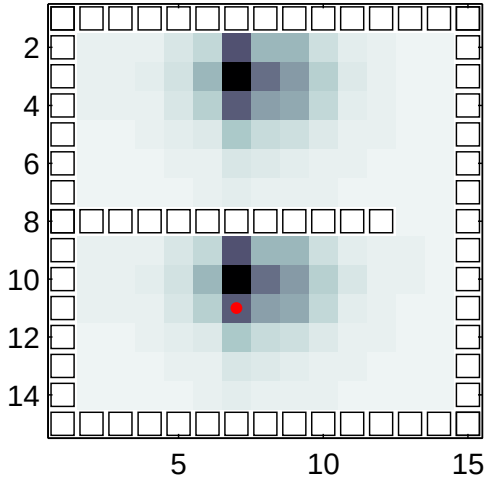

2
4
6
8
10
12
14
5
10
15

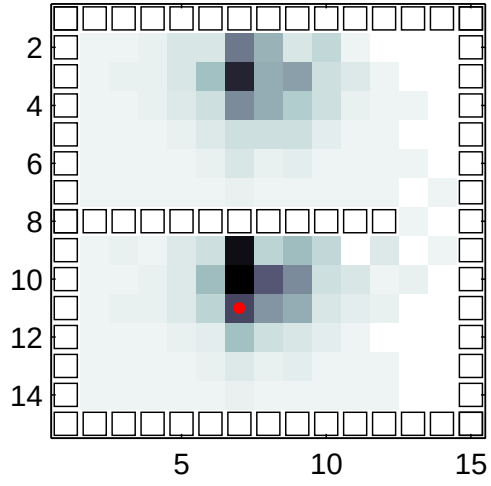

2
4
6
8
10
12
14
5
10
15

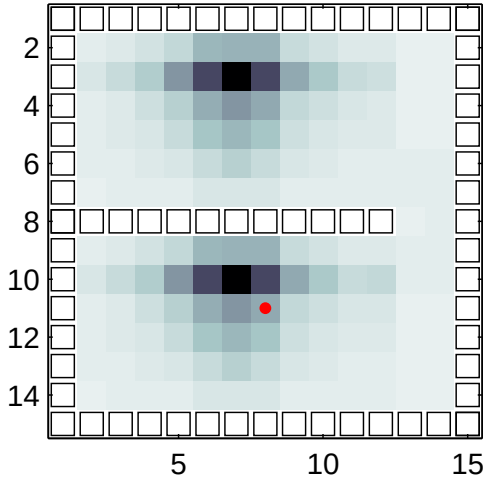

2
4
6
8
10
12
14
5
10
15

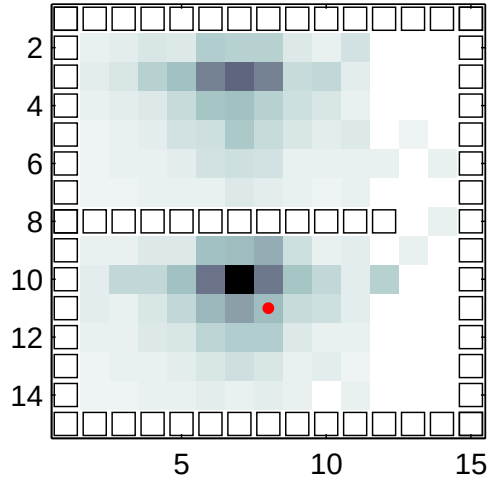

2
4
6
8
10
12
14
5
10
15

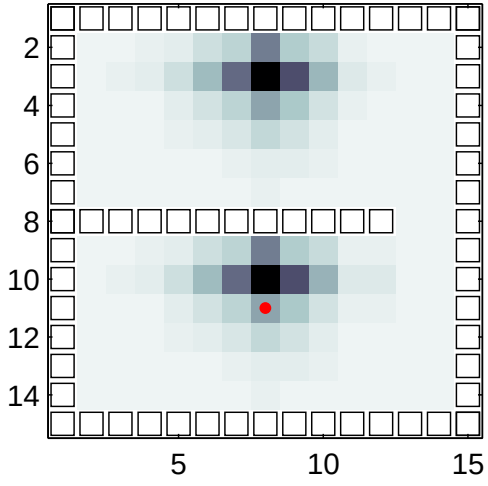

2
4
6
8
10
12
14
5
10
15

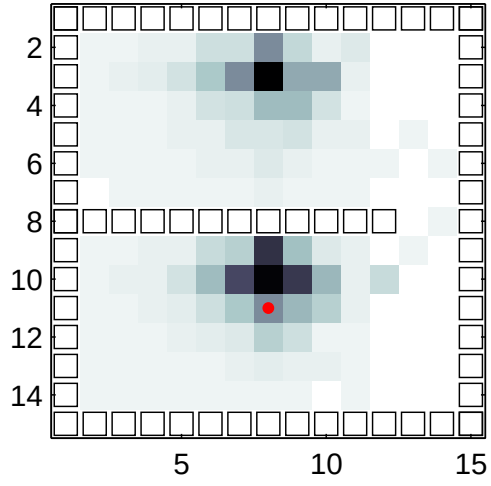

2
4
6
8
10
12
14
5
10
15

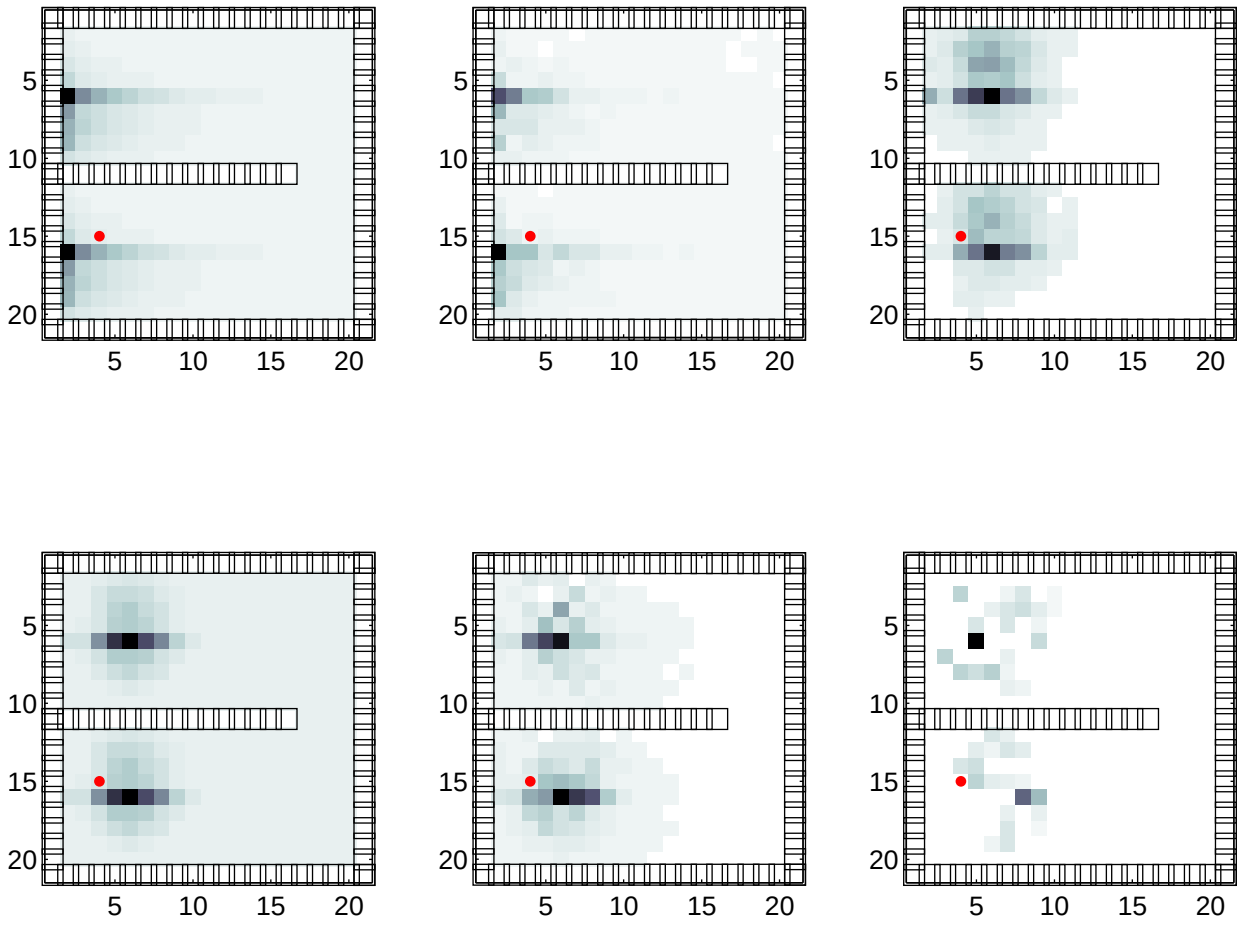

2
4
6
8
10
12
14
5
10
15

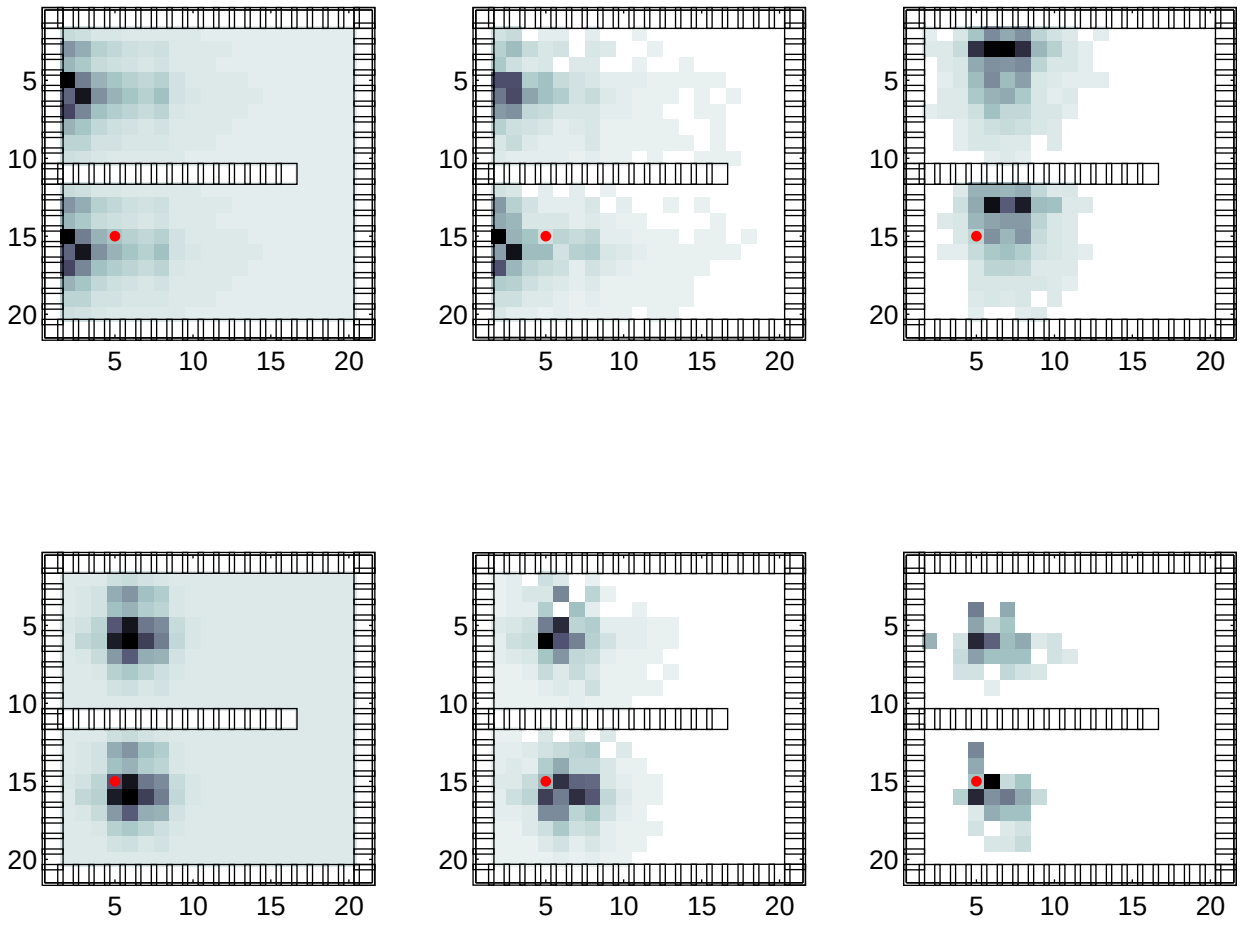

2
4
6
8
10
12
14
5
10
15

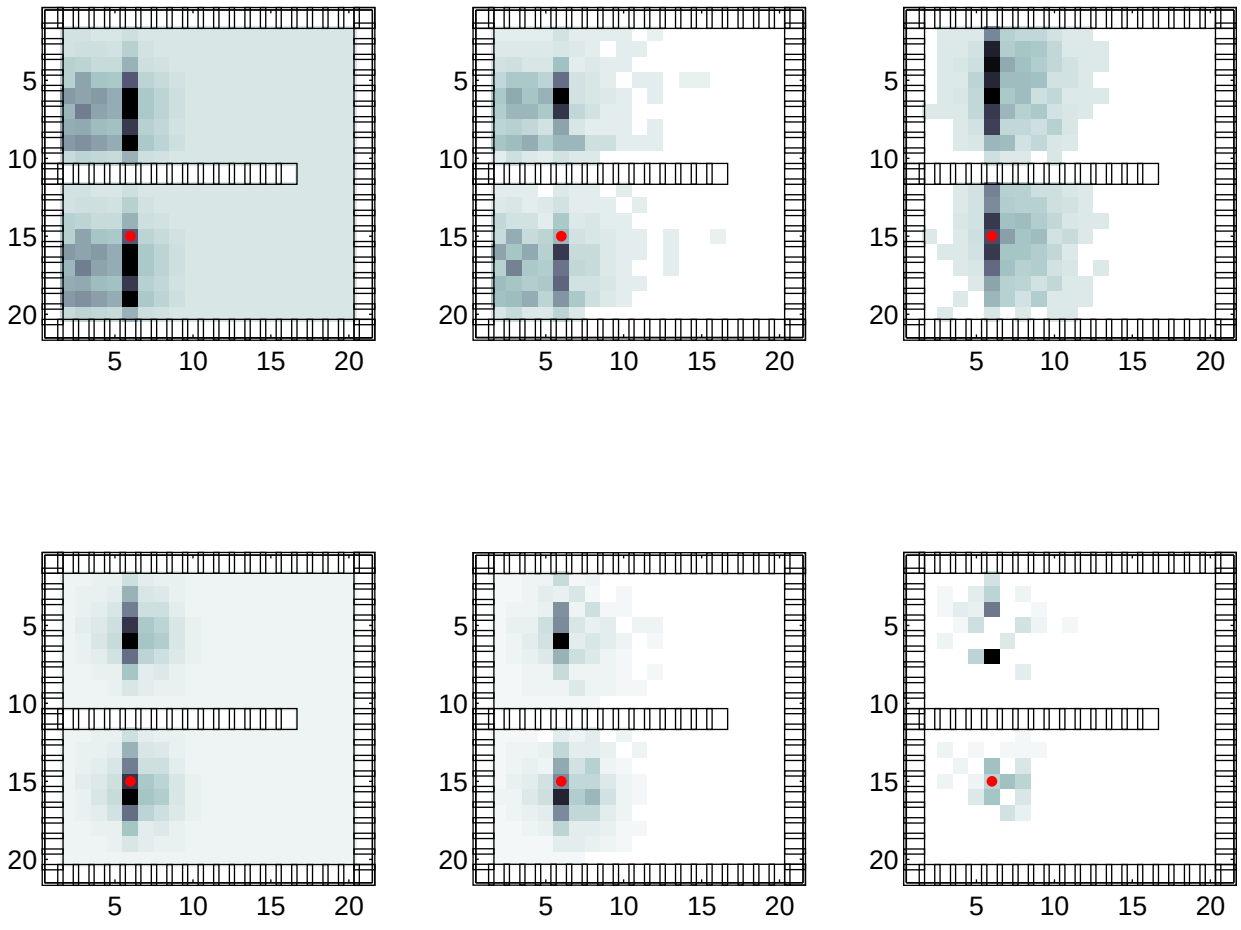

2
4
6
8
10
12
14
5
10
15

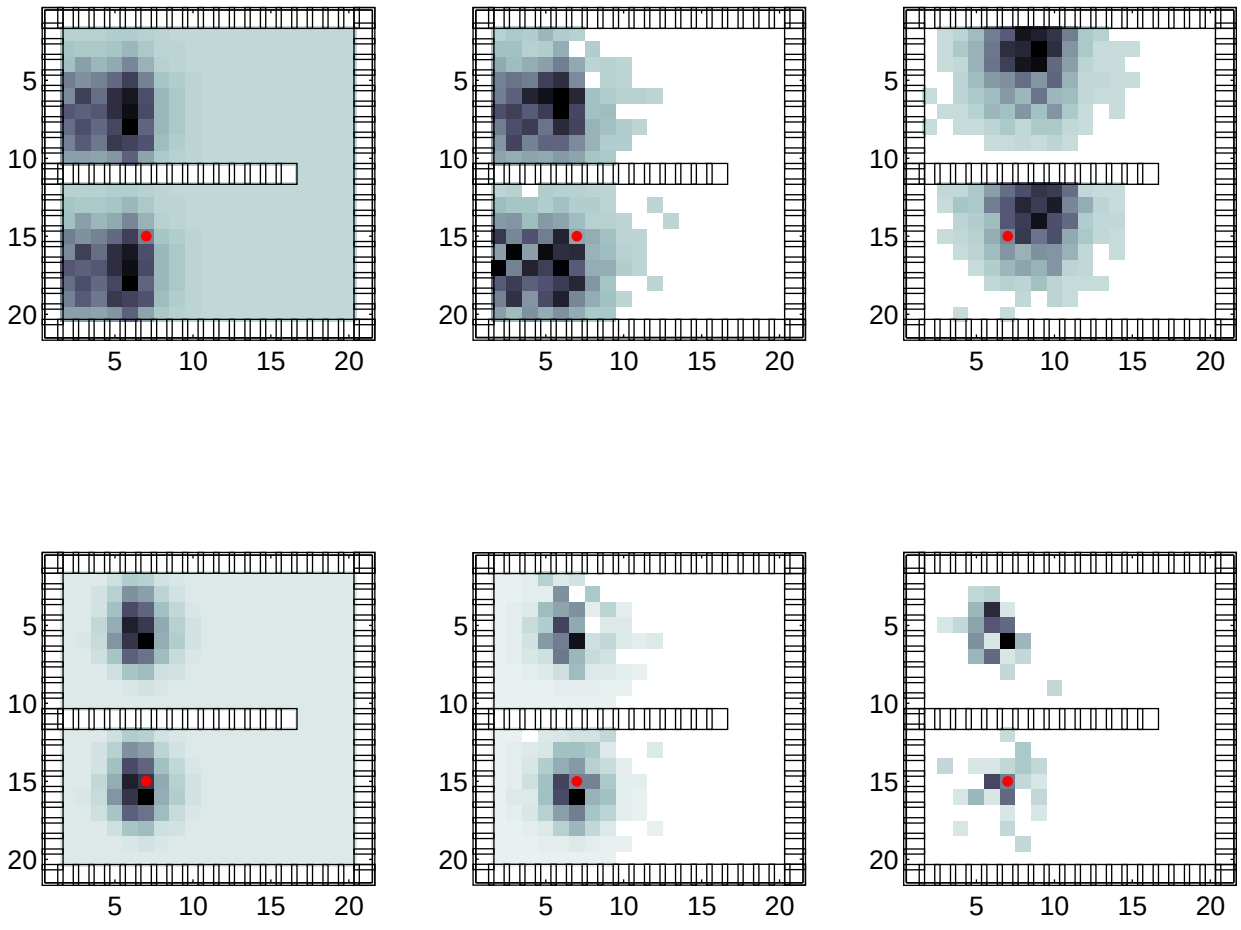

2
4
6
8
10
12
14
5
10
15

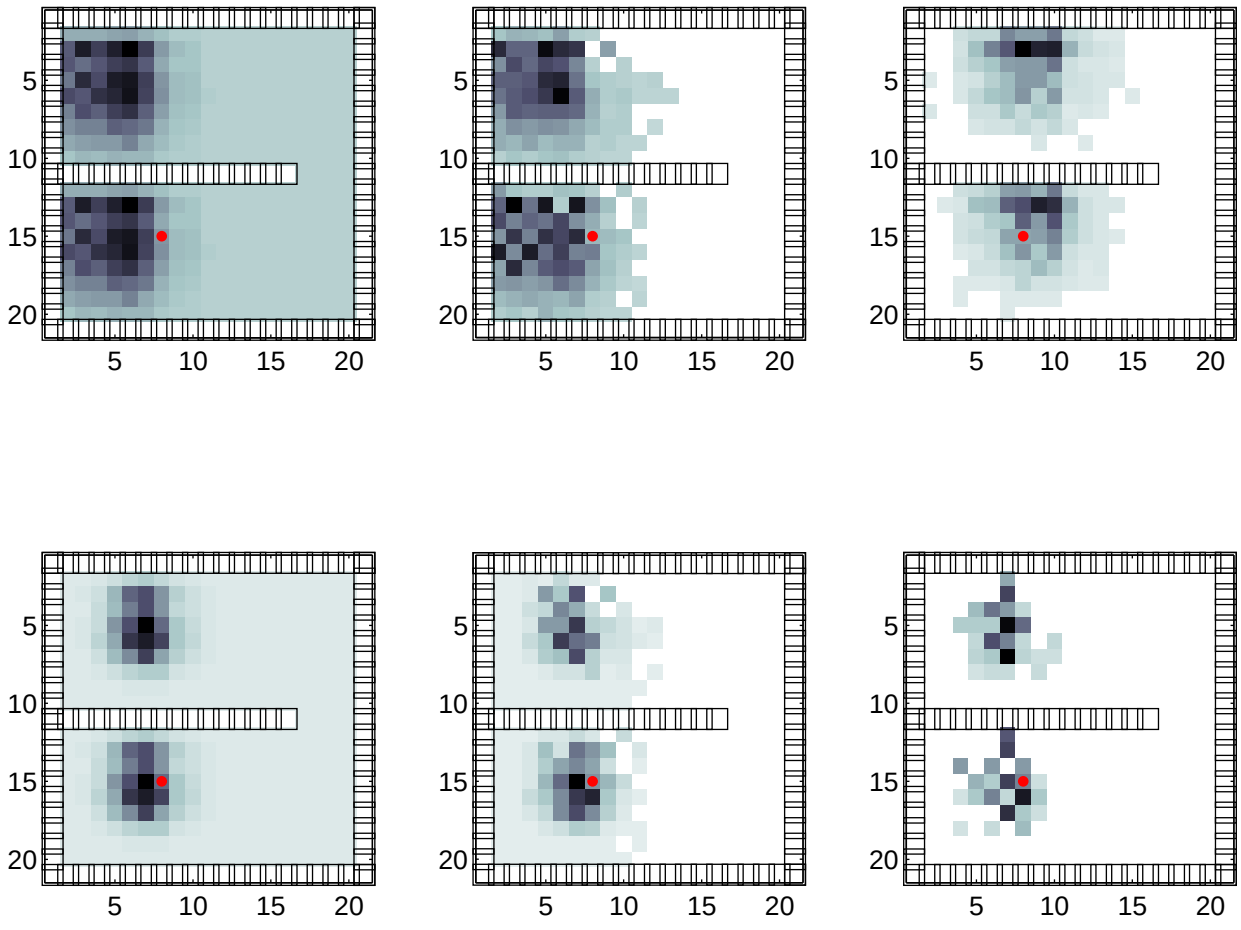

2
4
6
8
10
12
14
5
10
15

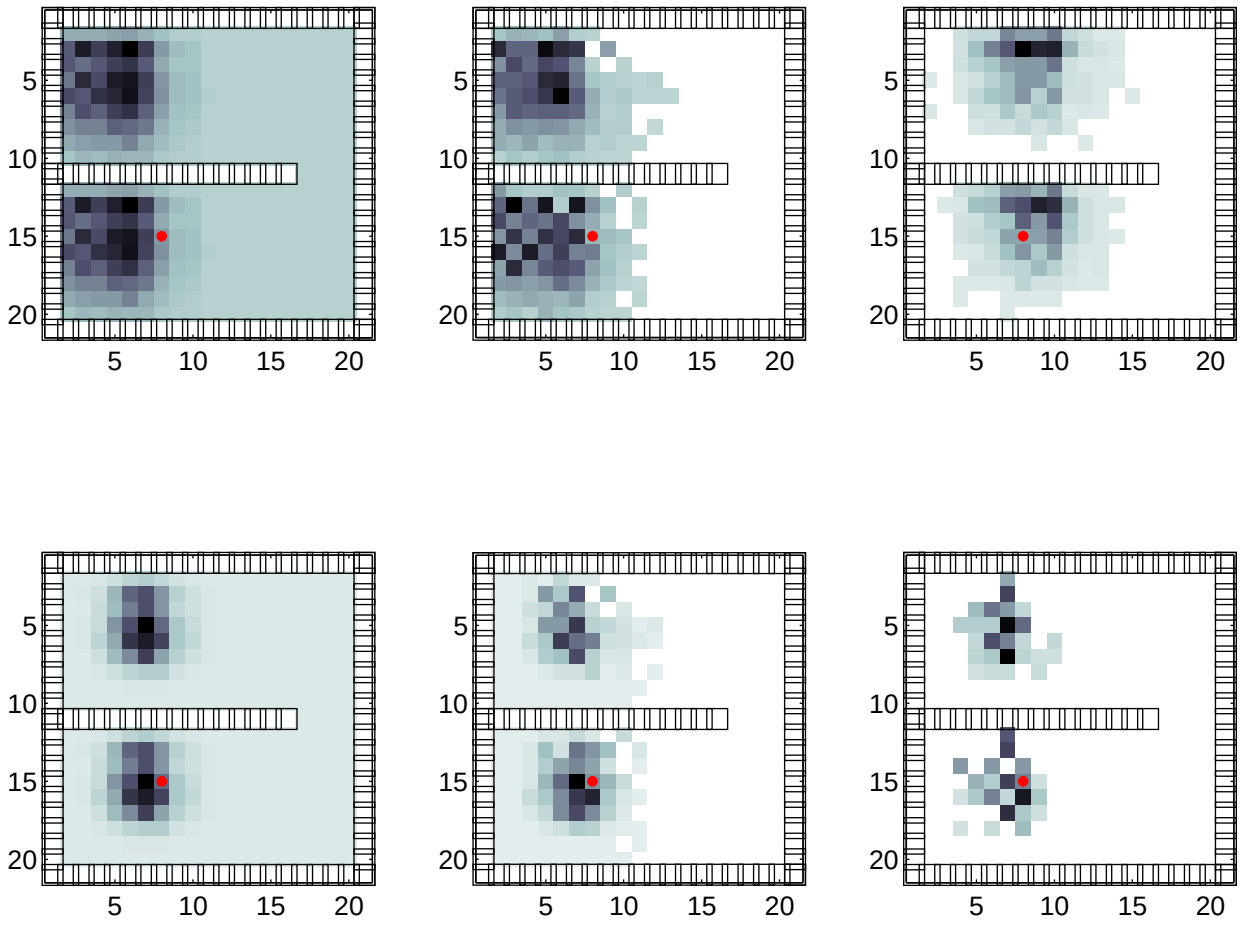

5
10
15
20
5
10
15
20

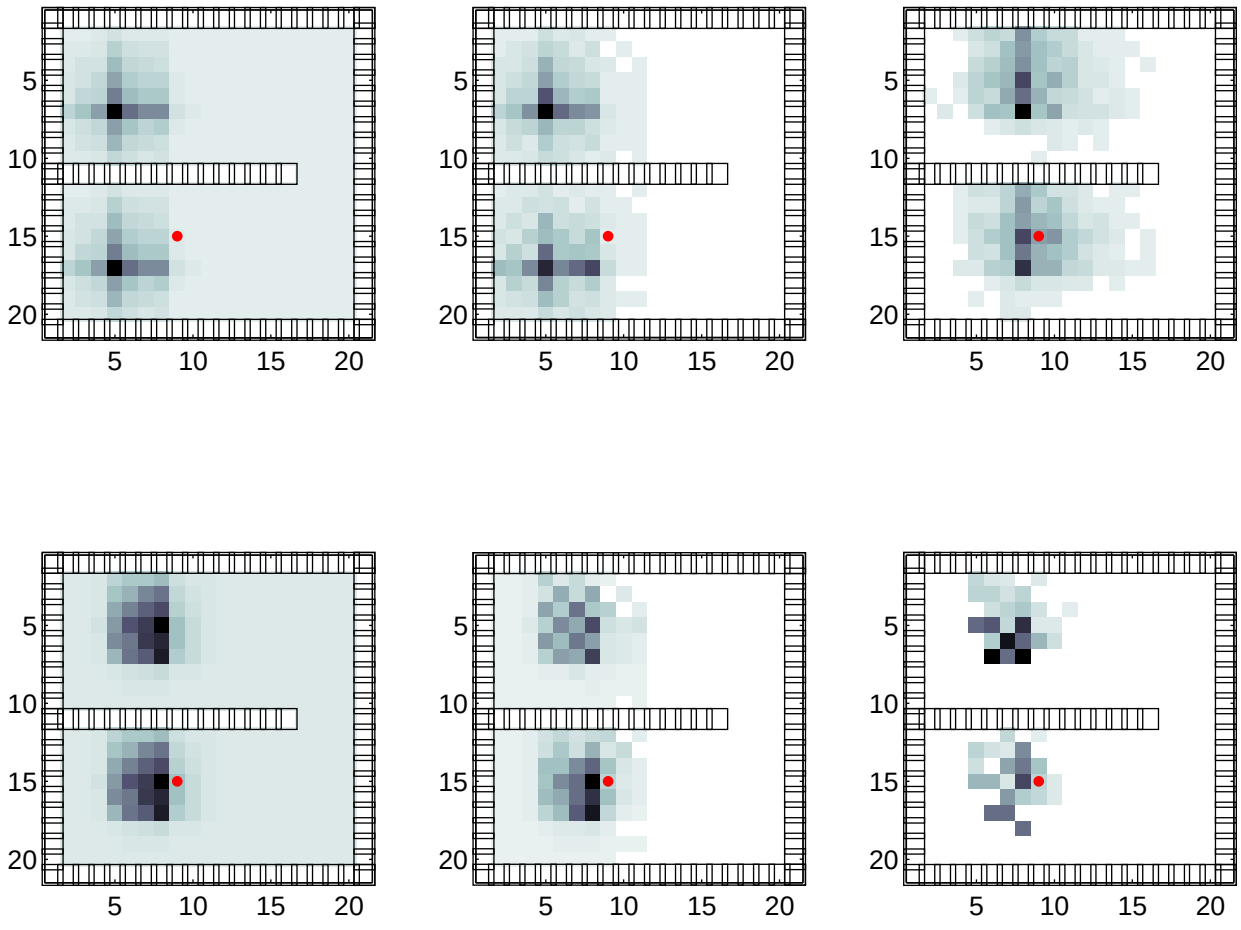

5
10
15
20
5
10
15
20

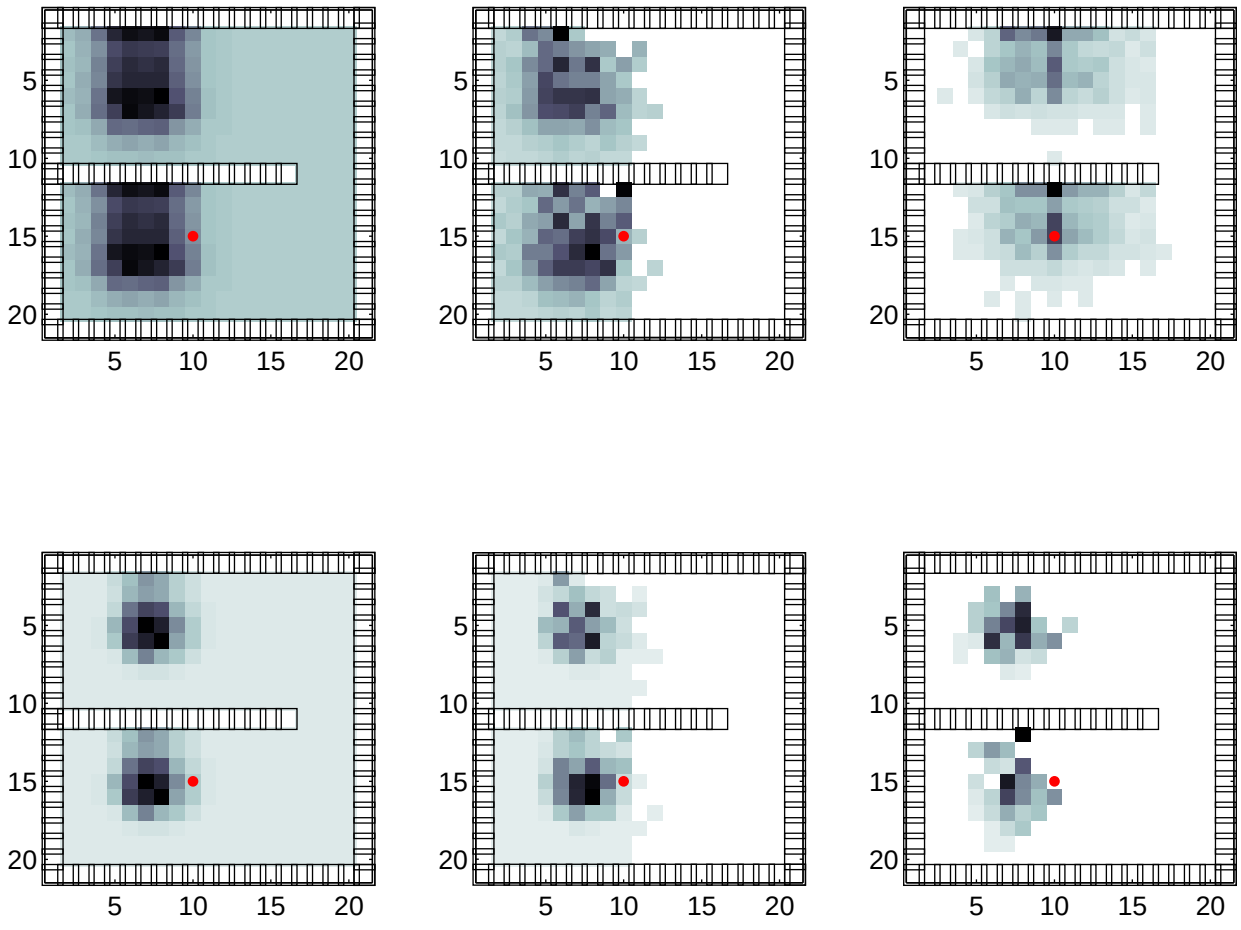

5
10
15
20
5
10
15
20

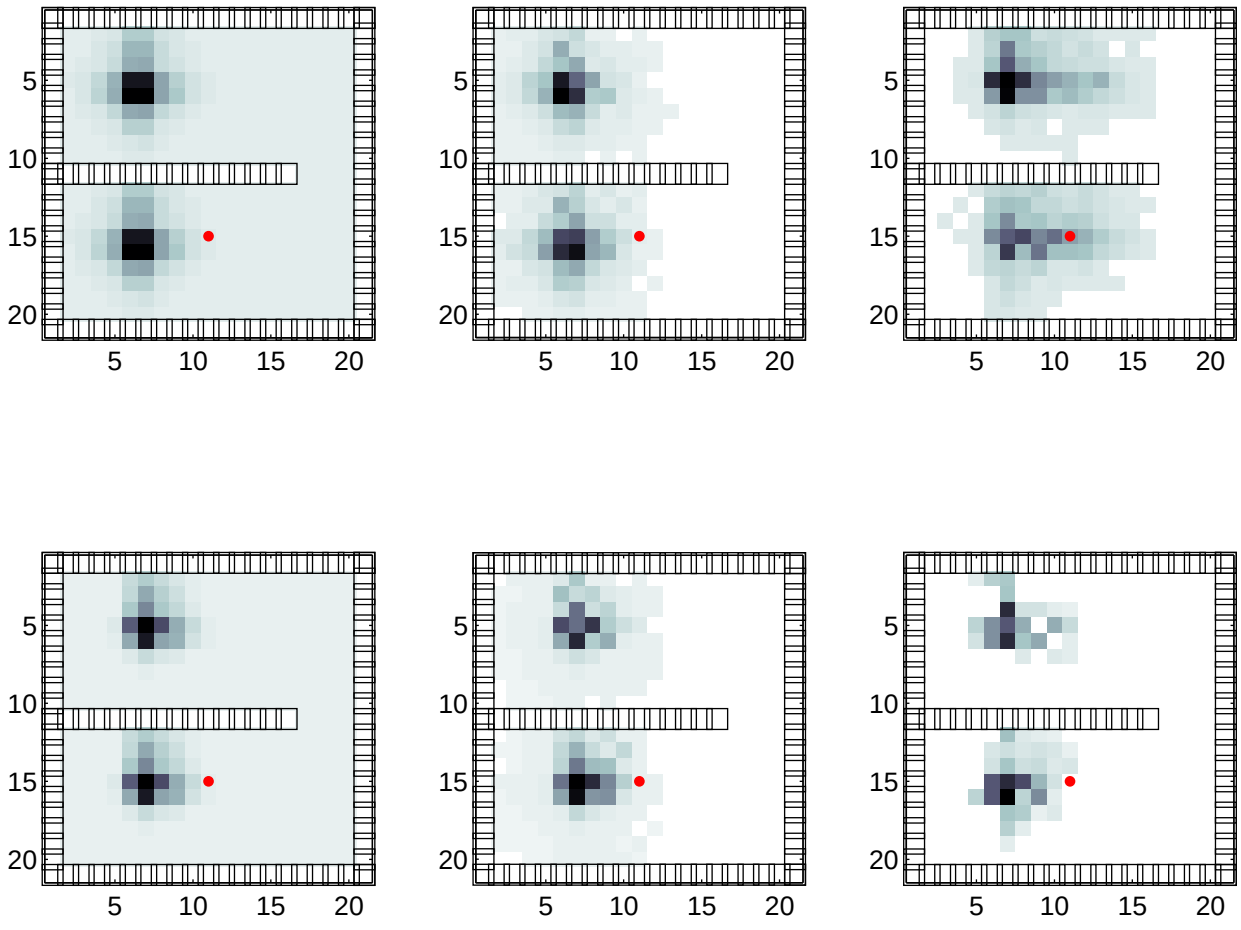

5
10
15
20
5
10
15
20

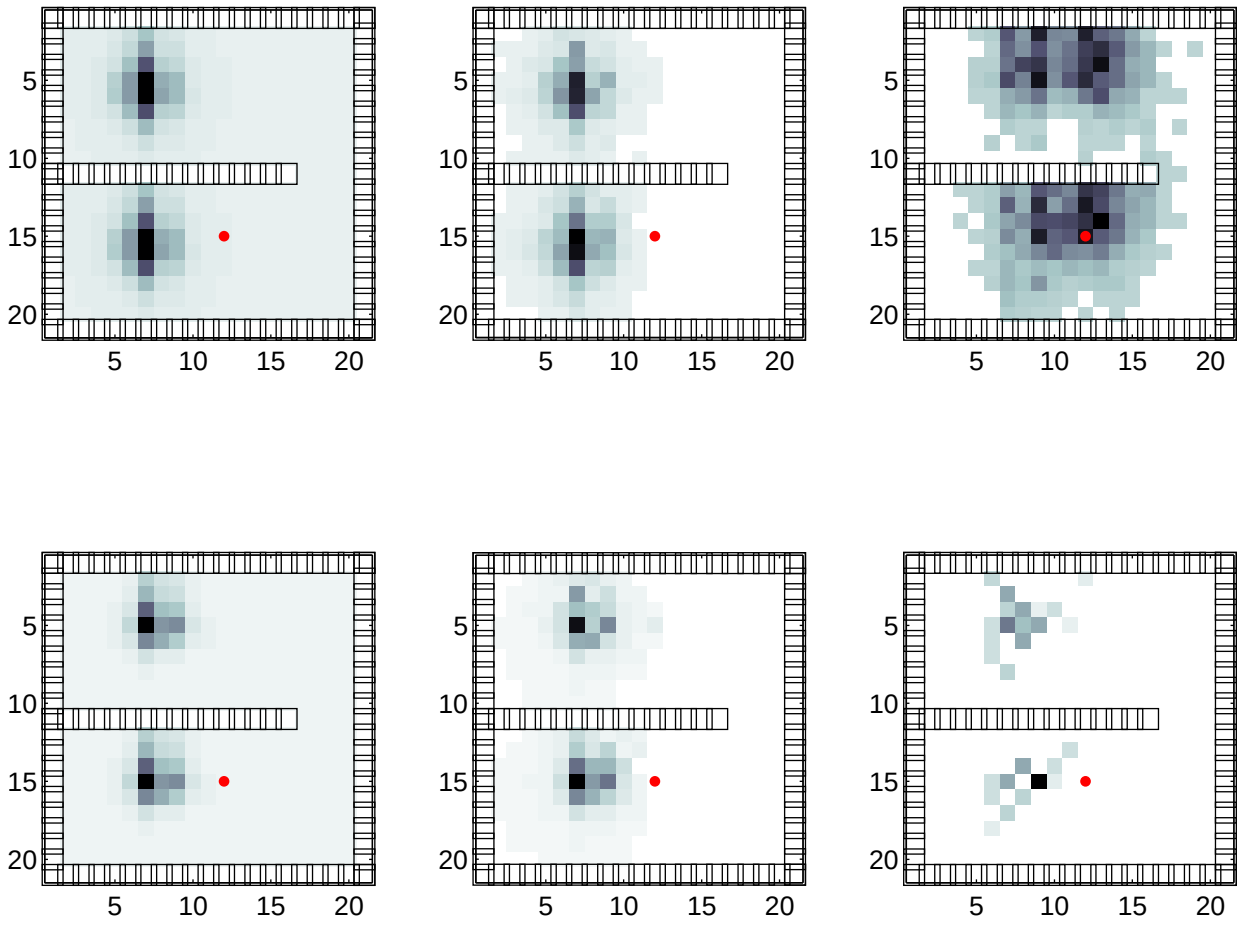

5
10
15
20
5
10
15
20


5
10
15
20
5
10
15
20


5
10
15
20
5
10
15
20


5
10
15
20
5
10
15
20


5
10
15
20
5
10
15
20


5
10
15
20
5
10
15
20

Conclusions

- Theory behind FFBS is now fully understood. Contrarily to what was thought, time-uniform asymptotics are preserved for mixing kernels.
- Two main problems remain:
  1. The complexity is generally high (grows as the square of the number of particles, though significant complexity reduction can be achieved in general): the two-filter algorithm is a solution.
  2. The support of the smoothing distribution is the same than the support of the filtering distribution: better algorithms can be obtained using either pilot sampling or iterative filtering.